Of Stacks and Muses: Adventures in Learning Analytics at Marist College

Eitel Lauría, Edward Presutti, Maria Kapogiannis

Of Stacks and Muses: Adventures in Learning Analytics at Marist College

Eitel J. M. Lauría; Edward Presutti; Maria Kapogiannis
Marist College, Poughkeepsie, NY, USA

LXAI @ ICML 2019
Long Beach, CA
June 10, 2019

Abstract

Marist Universal Student Experience (aka MUSE) is a predictive modeling application based on machine learning techniques that provides early alert and detection of academically at-risk students. The system uses data from previous semesters enriched with student performance and demographics data to train classifiers that make predictions of student academic performance a few weeks into the semester. MUSE implements a stacked ensemble architecture to build early detection models. This machine learning method is mostly absent and minimally referenced in the learning analytics literature. Hence, this work makes two relevant contributions:1) it describes a methodology for building a stacked ensemble architecture learnt from data; 2) it provides proof of concept of how stacked ensembles can be applied in the context of early detection of academically at-risk students. Experimental tests are carried out to demonstrate the predictive performance of the stack when making predictions on college-wide data.

Motivation

The numbers are appalling: in the US, the average six year degree completion below 60% for public 4-year institutions (slightly above 65% for private), with percentages plummeting when considering black student populations or Hispanic student populations. Four-year graduation rates are even more worrisome (see Table1 below).

Table 1. Average Graduation Rates in the US

	4-year graduation rate		6-year graduation rate	
Public 4-year colleges and universities	All	33.3%	All	57.6%
	– Black	17.4%	– Black	40.3%
	– Hispanic	23.8%	– Hispanic	50.6%
Private non-for-profit 4-year colleges and universities	All	52.8%	All	65.4%
	– Black	29.7%	– Black	44.7%
	– Hispanic	46.8%	– Hispanic	61.0%

Source: College Completion, The Chronicle of Higher Education https://collegecompletion.chronicle.com/
Data include 3,800 degree-granting institutions in the US that reported a first-time, full-time degree-seeking undergraduate cohort, had a total of at least 100 students at the undergrad. level in 2013, and awarded undergrad. degrees between 2011 and 2013.

MUSE: a little bit of history

Open Academic Analytics Initiative (2011-2013)	• EDUCAUSE NGLC grant, funded by the Gates Foundation. • Create an early alert framework of academically at-risk students based on open source tools. • Pilots at community colleges and HBCUs
LAP v1 (2013-2017)	• Became part of Apereo's learning analytics initiative (Learning Analytics Processor). • Chosen in 2015 as a key component of the UK's national analytics infrastructure.
LAP v2 aka MUSE (2017-2019)	• Engine based on a stacked ensemble. • Web-based dashboards, integrated into LMS (Sakai). • Open source contribution to Apereo.

How does MUSE work?

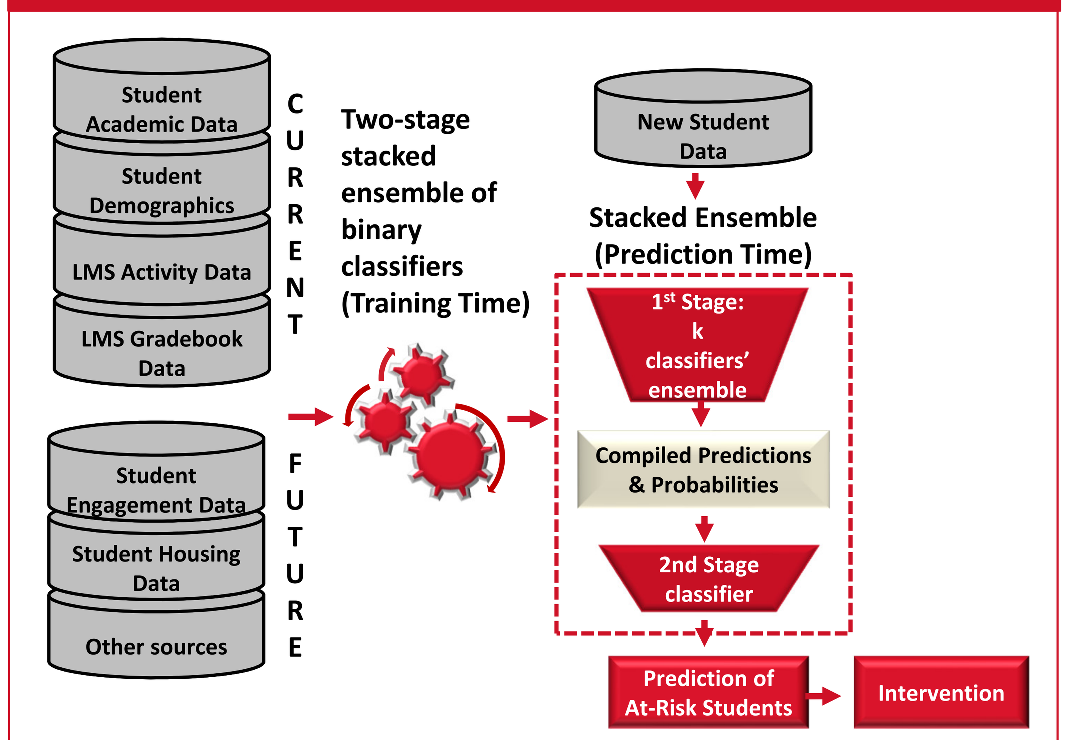


Figure 1. MUSE Architecture

Table 2. MUSE input data features

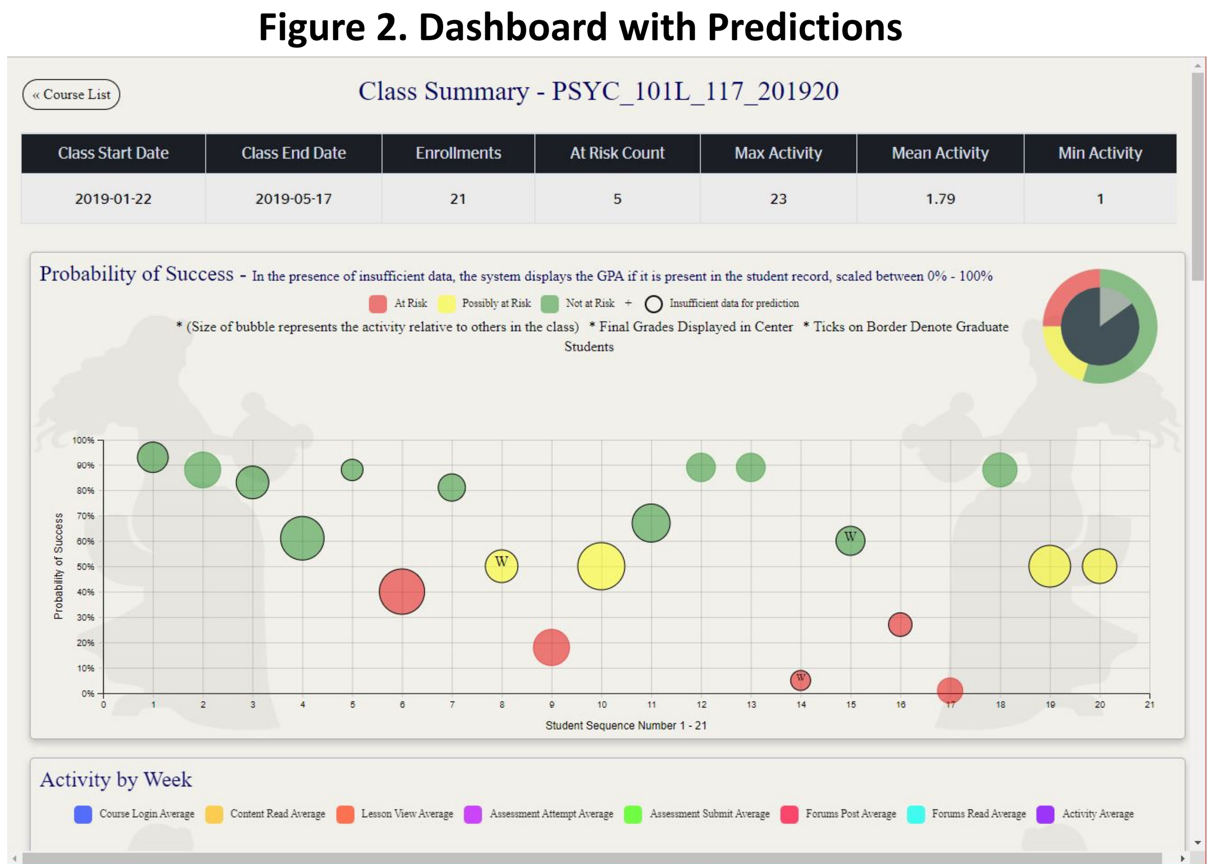
Predictors
Gender, Age, Class (Freshman, Sophomore, Jr, Sr.), Aptitude Score (e.g. SAT), Cumulative GPA, Course size, Discipline (SCI, LA, CSM, BUS, SBS, CA), LMS Total Activity (weeks 1-6 + sum), LMS login (weeks 1-6 + sum), LMS Content Read (weeks 1-6 + sum), Gradebook Composite Score (weeks 1-6)
Target feature: Academic Risk (Yes=at risk; No=good standing)

MUSE Dashboards

MUSE provides predictions of academic success of undergraduate students in a given course, six weeks into the semester of a 15-week course. The threshold of good academic standing is a letter grade C (students with less than a C are considered at risk). To make the predictions less crisp, we tie these predictions to a probability value based on the aforementioned threshold, and we use a color coding (see Figure 2).

- **GREEN** for students in good standing (those with a probability of success > 55%)
- **YELLOW** for student with an undetermined risk status (probability of success between 55% and 45%)
- **RED** for at-risk students (those with a probability of success < 45%).

Figure 2. Dashboard with Predictions



Stacked ensemble of classifiers

Training and testing a two-stage stack with *k* binary classifiers in the first stage, one binary classifier in the second stage, and 3 independent data sets A, B, C, to avoid data leakage (see Figure 3). After the stack is trained, tuned and tested, it can be used to make predictions on new data D. Figure 4 depicts the two-stage stack making predictions on incoming (and therefore unlabeled) data **D.**

Figure 3. Training and testing a two-stage stack

$$\mathbf{A}[X;y] \rightarrow \overbrace{\mathbf{A}[X;y] \underset{cv}{\leftarrow} \mathcal{C}_1 \ \ldots \ \mathbf{A}[X;y] \underset{cv}{\leftarrow} \mathcal{C}_k}^{\text{Train}}$$

$$\text{(i) Stage 1: Training} \quad \mathcal{M}_1 \quad \ldots \quad \mathcal{M}_k$$

$$\mathbf{B}[X;y] \rightarrow \overbrace{\mathbf{B}[X;y] \leftarrow \mathcal{M}_1 \ \ldots \ \mathbf{B}[X;y] \leftarrow \mathcal{M}_k}^{\text{Test}}$$

$$\text{ii) Stage 1: Testing} \quad \text{Compute } \hat{Y}_B;\hat{P}_B: \quad (\hat{y}_1^B;\hat{p}_1^B) \quad \ldots \quad (\hat{y}_k^B;\hat{p}_k^B)$$

$$\text{iii) Augment } \mathbf{B}: \quad \mathbf{B}[X;y] \rightarrow \mathbf{B}^{(aug)}[X;\hat{Y}_B;\hat{P}_B;y]$$

$$\mathbf{B}^{(aug)}[X;\hat{Y}_B;\hat{P}_B;y] \rightarrow \overbrace{\mathbf{B}^{(aug)}[X;\hat{Y}_B;\hat{P}_B;y] \underset{cv}{\leftarrow} \mathcal{C}_{SS}}^{\text{Train}}$$

$$\text{(iv) Stage 2: Training} \quad \mathcal{M}_{SS}$$

$$\mathbf{C}[X;y] \rightarrow \overbrace{\mathbf{C}[X;y] \leftarrow \mathcal{M}_1 \ \ldots \mathbf{C}[X;y] \leftarrow \mathcal{M}_k}^{\text{Predict (1st stage)}}$$

$$\text{(v) Compute } \hat{Y}_C;\hat{P}_C: \quad (\hat{y}_1^C;\hat{p}_1^C) \quad \ldots \quad (\hat{y}_k^C;\hat{p}_k^C)$$

$$\text{vi) Augment } \mathbf{C}: \quad \mathbf{C}[X;y] \rightarrow \mathbf{C}^{(aug)}[X;\hat{Y}_C;\hat{P}_C;y]$$

$$\mathbf{C}^{(aug)}[X;\hat{Y}_C;\hat{P}_C;y] \rightarrow \overbrace{\mathbf{C}^{(aug)}[X;\hat{Y}_C;\hat{P}_C;y] \leftarrow \mathcal{M}_{SS}}^{\text{Predict (2nd Stage)}}$$

$$\text{(vii) Stage 2: Testing} \quad (\hat{y}_{SS}^C;\hat{p}_{SS}^C)$$

Training and testing a stack with *k* classifiers in the first stage, 1 classifier in the second stage, and 3 independent datasets **A, B, C**.

Figure 4. Using the stack for prediction on new data

$$\mathbf{D}[X] \rightarrow \overbrace{\mathbf{D}[X] \leftarrow \mathcal{M}_1 \ \ldots \mathbf{D}[X] \leftarrow \mathcal{M}_k}^{\text{Predict (1st Stage)}}$$

$$\text{(I) Compute } \hat{Y}_D;\hat{P}_D: \quad (\hat{y}_1^D;\hat{p}_1^D) \quad \ldots \quad (\hat{y}_k^D;\hat{p}_k^D)$$

$$\text{(II) Augment } \mathbf{D}: \quad \mathbf{D}[X] \rightarrow \mathbf{D}^{(aug)}[X;\hat{Y}_D;\hat{P}_D]$$

$$\mathbf{D}^{(aug)}[X;\hat{Y}_D;\hat{P}_D] \rightarrow \overbrace{\mathbf{D}^{(aug)}[X;\hat{Y}_D;\hat{P}_D] \leftarrow \mathcal{M}_{SS}}^{\text{Predict (2nd Stage)}}$$

$$\text{(III) Make stack predictions} \quad (\hat{y}_{SS}^D;\hat{p}_{SS}^D)$$

After the stack is trained, tuned and tested, it can be used to make predictions on new data **D**. Learning algorithms used to train base classifiers should be as different as possible so that the predictions made by them have relatively low correlations (< 0.75 - 0.80). If predictions are highly correlated, it indicates that the base models map very similar hypothesis functions, which defeats the purpose of using a stack.

Experimental Setup

We studied the use of a two-stage stack -which drives MUSE- trained with undergraduate data from 10 semesters (Fall 2012 – Spring 2017). LMS student activity data is recorded as weekly frequency ratios, normalized with the mean and std dev. of each course. The target variable is the final grade, recoded as a binary variable –Academic Risk- using letter grade C as threshold (see Table 2). A random sample of 31029 records (35% of the total) was used. We performed 8 batches of experiments, using 2 different configurations of classification algorithms for the first-stage (base) models; 2 different sets of predictors to train the base models; and 2 different algorithms for the second-stage model. Each batch was repeated ten times with varying random generator seeds to account for variation in predictive performance due to the data; in each run the data was randomly partitioned into datasets A, B and C with 10343 records each. This amounted to a total of 80 runs in the experiment (2 x 2 x 2 x 10). See Tables 3-5 for details.

Table 3. First-stage Classifiers

Code	Description
RF	The Random Forests algorithm (Breiman, 2001), a variation of bagging applied to decision trees
NN	A feed-forward neural network (multilayer perceptron) with one hidden layer and varying number of units.
NB	The naïve Bayes algorithm with kernel estimation, to estimate the densities of numeric predictors.
XB	XGBtree (Chen and Guestrin, 2016), a recent implementation of the gradient boosted tree algorithm.

Table 4. Second-stage Classifiers

Code	Description
LOG	Regularized logistic regression using the the LibLinear library (Fan et al., 2008)
LMT	Logistic model trees (Landwehr et al., 2005)

Table 5. First-stage Predictor Configuration

Code	Description
ALL	All predictors (as described in Table 2)
NoCS	All predictors except the Gradebook composite score

Results and Discussion

Table 6 displays the assessment of mean predictive performance (mean AUC) of the stacked ensemble for the eight experiments described. The best performing 1st stage classifier's mean AUC is reported for comparison purposes The stack exhibited very good predictive performance when trained with all first stage predictors, outperforming all three base classifiers for both configurations of first base classifiers (XB+NN+RF and XB+NN+NB). For the XB+NN+RF/LMT stack, the mean AUC value was 0.935; for the XB+NN+RF/LOG stack, the mean AUC value was 0.939. The naïve Bayes (NB) algorithm was considerably less performant than the random forests algorithm -compare mean AUC(NB)=0.858 with mean AUC(RF)=0.920-, but all three classification algorithms in the XB+NN+NB configuration are considerably different from each other, as exhibited by the correlations between predicted probabilities (see Table 7). The absence of Gradebook data had an expected negative impact on the predictive performance of the stack, reducing its average AUC to 0.855 for the XB+NN+RF configuration, and to values of 0.851 and 0.852 for the XB+NN+NB configuration. The stack's predictive performance remained superior on average, and significant differences between mean AUC values of the stack and its component classifiers were present in all but one configuration (XB+NN+NB/LMT). Still, it is noteworthy how well the stack performed despite the relatively weaker performance of the naïve Bayes classifier. This reveals another advantage of the stacked ensemble architecture: it cushions weaker performances of its components, promoting more stable predictions when faced with varying characteristics of the data.

Table 6. Stack Predictive Performance Results

Stage 1 Classifiers	Stage 1 Predictors	Stage 2 Classif.	Stack AUC		BPS1C AUC
			Mean	SE	
XB+NN+RF	ALL	LMT	0.934	0.003	0.928
XB+NN+RF	ALL	LOG	0.936	0.002	0.928
XB+NN+RF	NoCS	LMT	0.855	0.001	0.846
XB+NN+RF	NoCS	LOG	0.855	0.001	0.846
XB+NN+NB	ALL	LMT	0.933	0.002	0.928
XB+NN+NB	ALL	LOG	0.933	0.002	0.928
XB+NN+NB	NoCS	LMT	0.851	0.002	0.846
XB+NN+NB	NoCS	LOG	0.852	0.001	0.846

Table 7. Correlations of Predicted Probabilities

	Mean	Std Dev	Min	Max
XB-NN	0.77	0.11	0.44	0.92
NN-RF	0.76	0.11	0.49	0.89
RF-XB	0.85	0.07	0.7	0.96
NN-NB	0.22	0.06	0.09	0.33
NB-XB	0.34	0.06	0.24	0.45

Conclusion and Limitations

- Current choice of classifiers is discretionary (state of the art classifiers that yield probabilities and different enough to cover the hypothesis space).
- MUSE uses a fixed 2-stage stack with 3 base classifiers.
- Nonetheless, it provides first-time insight of the use of a stacked ensemble architecture in the domain of learning analytics and early detection of academically at-risk students.

Corresponding Author:
Eitel J.M. Lauría, PhD
Professor, School of Computer Science & Mathematics
Marist College
Poughkeepsie, NY, USA
Eitel.Lauria@marist.edu
+1 845 575 3598

Relevant References:

- Chen T., Guestrin C., 2016. XGBoost: A Scalable Tree Boosting System. CoRR abs/1603.02754.
- Jayaprakash S., Moody E., Lauría, E., Regan, J., Baron, J., 2014. Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative. J. Learn. Anal. 1, 42.
- Lauría E., Presutti E., Sokoloff M., Guarino M., 2016. Crossing the Chasm to Big Data: Early Detection of at-Risk Students in a Cluster Computing Environment. Presented at the Proceedings of the 7th International Learning Analytics & Knowledge Conference (LAK'17)- Practitioner Track. Vancouver, Canada.
- Lauría E., Presutti E., Kapogiannis M., Kamath A., "Stacking Classifiers for Early Detection of Students at Risk", Proceedings of the 10th International Conference of Computer Supported Education (CSEDU 2018), Funchal Madeira, Portugal, March 15-17, 2018, vol 1, 390-397
- Wolpert D., 1992. Stacked generalization. Neural Netw. 5, 241–259.